Stein's Method

Matthew Kwan

March 31, 2014

Contents

Comment 0.1. Here's the general plan:

- Introduction: introduce with limit theorems (weak convergence) to motivate Stein's method
- probability theory review
- discussion of weak convergence, definition of stein metrics and proofs that stein metrics are "compatible" with weak convergence.
  - Highlight: proof that bounded lipschitz metric metrizes the weak topology
- general discussion of the idea of stein's method
  - bare-hands proof of Berry-Esseen theorem
- Exchangeable pairs
  - fixed points in permutations example
- (maybe) Size-bias coupling
  - subgraphs of erdos-renyi random graphs

1 Introduction

The *central limit theorem*, roughly speaking, says that sums of independent random variables are "approximately normal", with the approximation improving as the sumber of terms in the sum increases. Similarly, the *Poisson limit theorem* says that the number of occurences of independent "rare" events over a given time period is "approximately Poisson". These two limit theorems are archetypal examples of a large class of related results in statistics and probabilistic combinatorics.

These types of results are generally formally stated asymptotically, with a particular type of convergence called *weak convergence*, or *convergence in distribution*. Such results are very powerful, but an obvious shortcoming is that they do not quantify the convergence any way.

It may be that a sequence of random variables is “asymptotically normal”, but we cannot say that any particular random variable in that sequence is individually “close to normal”. Even if we are only interested in asymptotic results, it can be problematic that we cannot say convergence is “uniform” in any sense.

The solution to both these problems is to define a distance metric between probability distributions, and study convergence with respect to this metric. Fortunately, weak convergence is in fact convergence with respect to a topology, and this topology is metrizable. That is, there is a metric on the set of probability distributions such that the convergence in this metric space is equivalent to weak convergence.

Stein's method was introduced by Charles Stein in 1972. It is most generally a method for approximating expected values. However, at least in probabilistic combinatorics, Stein's method has proved especially useful for bounding the distance between probability distributions, in a variety of metrics consistent with weak convergence. This can be used to quantify existing limit theorems, and can also be used as a tool to prove purely asymptotic results.

In this thesis, we set out the theoretical groundwork for Stein's method, including a rigorous overview of probability theory and a discussion of weak convergence. We then present a very general framework for the Stein's method, and outline a few specific ways of applying the method. We include a number of specific examples and applications.

Unless stated otherwise, all proofs are “original” in that I came up with them, although many are quite straightforward and probably exist elsewhere. For the proofs that I have adapted from other sources, I tried to add some new explanation or details.

1.1 Notation and Assumed Knowledge

For this thesis, the set of natural numbers $\mathbb{N}$ includes zero. We write $\mathbf{1}_A$ for the characteristic function of a set A: $\mathbf{1}_A(x) = 1$ if $x \in A$, otherwise $\mathbf{1}_A(x) = 0$. The function f restricted to the set A is denoted $f|_A$. The falling factorial $n(n-1)\dots(n-k+1)$ is denoted $(n)_k$. Finally, $[k]$ denotes the set $\{1, \dots, k\}$.

Unless otherwise specified, all asympotics are as $n \to \infty$. Apart from standard asymptotic notation, $f \sim g$ means $f = g(1 + o(1))$, $f \prec g$ means $f = O(g)$ and $f \asymp g$ means $f = \Theta(g)$.

In this thesis, unless stated otherwise, graphs are labelled. That is, they are distinguished even within isomorphism classes. A graph may not have loops or multiple edges; an object which is allowed to have loops and/or multiple edges will be called a multigraph. We write $G \subseteq G'$ to indicate that G is a (not necessarily induced) subgraph of G'.

I will occasionally use results from analysis without proof. I will usually refer to a numbered theorem in Rudin's Real and Complex Analysis [Rud66] or Functional Analysis [Rud73] when doing so.

2 General Probability Theory

For many combinatorial applications, an informal understanding of probability theory (often considering only discrete spaces) will suffice. However, in this thesis a rigorous foundation in probability theory will be useful. This section will therefore assume knowledge of basic measure theory; see, for example, [Rud66]. However, where possible, we will note any simplifications that arise from assuming discreteness, for the benefit of those less familiar with general probability theory.

No knowledge of probability theory is assumed; the first few subsections will briefly review the foundations of probability. The reader may nevertheless want to refer to a probability theory book such as [Kal02] for some additional detail and further reading.

2.1 Probability Spaces

Definition 2.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity.

Remark 2.2. For our purposes Ω will often be countable, with $\mathcal{A}$ as the power set of Ω. In this case $\mathbb{P}$ is uniquely defined by the probabilities $\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, for each $\omega \in \Omega$. We will discuss specific probability spaces on combinatorial objects in Section 3.

Definition 2.3. An *event* in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a measurable set $A \in \mathcal{A}$.

For an event A, $\mathbb{P}(A)$ is interpreted as the "probability that A occurs". Events will usually be of the form $A = \{\omega \in \Omega : P(\omega) \text{ holds}\}$, where $P(\omega)$ is some property of an object ω. For clarity, we often abuse notation slightly and write $\mathbb{P}(P \text{ holds})$ instead of $\mathbb{P}(\{\omega \in \Omega : P(\omega) \text{ holds}\})$.

2.2 Random Elements

Definition 2.4. A *random element* is a measurable function $X : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ between measurable spaces. If $\Omega_2 = \mathbb{R}^n$ for some $n \in \mathbb{N}$, with $\mathcal{A}_2$ the Borel σ-algebra on $\mathbb{R}^n$, then we say X is a *random vector.* A one-dimensional random vector is a *random variable.* If Ω_2 is countable then we say X is *discrete.*

Remark 2.5. Especially in combinatorial spaces, Ω_1 is often countable. In this case, any function from a probability space $\big(\Omega_1, 2^{\Omega_1}, \mathbb{P}\big)$ is measurable.

To interpret a random variable, we need a probability measure $\mathbb{P}$ on the underlying measurable space $(\Omega_1, \mathcal{A}_1)$ (often, this will be implicit). Then, $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega_1 : X(\omega) \in A\})$ is the probability that X takes a value in the set A. Often, we will only be interested in such probabilities: that is, we do not care about the realization of a random variable as a function on an underlying probability space. This motivates the following definition:

Definition 2.6. Suppose X is a random element which takes values in the measurable space $(\Omega, \mathcal{A})$. The *distribution* (or *law*) $\mathcal{L}_X$ of X with respect to an underlying probability $\mathbb{P}$ is the pushforward measure with respect to X. That is, it is a probability measure defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for $A \subseteq \mathcal{A}$. Also, we occasionally use the notation $\mathcal{L}(X) := \mathcal{L}_X$ for ease of reading.

It is worth noting that in fact any probability measure is the distribution of some random element, so we can define a probability distribution in the abstract and then assert the existence of a random variable with that distribution. To see this, note that given a probability measure $\mathcal{L} \in \mathcal{P}(\Omega)$, we can choose $X = \mathrm{id}_\Omega$ to have $\mathcal{L}_X = \mathcal{L}$ with respect to the underlying probability measure $\mathcal{L}$. We also use the notation $X \in \mathcal{L}$ to indicate that X has distribution $\mathcal{L}$.

Example 2.7. The normal distribution with parameters μ and σ is denoted $\mathcal{N}(\mu, \sigma)$ or $\mathcal{N}_{\mu,\sigma}$ and is defined by $\mathcal{N}_{\mu,\sigma}(B) = \frac{1}{\sigma\sqrt{2\pi}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}\, x$ for any Borel set B.

Example 2.8. The Poisson distribution with parameter λ is denoted $\mathrm{Po}(\lambda) = \mathrm{Po}_\lambda$; this is defined by $\mathrm{Po}_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for all $k \in \mathbb{N}$. (see Remark 2.2).

It is in general a little tricky to define “the set of values a random element can take”, but this is straightforward in the discrete case.

Definition 2.9. The *support* of a discrete random element X is the set

$$\mathrm{supp}(X) = \{k \in \Omega : \mathbb{P}(X = k) > 0\}.$$

2.3 Dependence and Coupling

Definition 2.10. Suppose X and X' are random elements $(\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$. We say that X and X' are *independent* if

$$\mathbb{P}(X \in A_2)\mathbb{P}\big(X' \in A_2\big) = \mathbb{P}\big(X \in A_2 \text{ and } X' \in A_2\big)$$

for all $A_2 \in \mathcal{A}_2$. If $A_1, A_1' \in \mathcal{A}$ then we analogously say A_1 and A_1' are independent if

$$\mathbb{P}(A_1)\mathbb{P}\big(A_1'\big) = \mathbb{P}\big(A_1 \cap A_1'\big).$$

We can similarly say an event is independent of a random element. If two objects are not independent, then we say they are *dependent*.

Intuitively, two objects are dependent if information about one object can give information about the other. For example, we might be interested in the probability of an event A, under the assumption that the event A' occurs.

Definition 2.11. The *conditional probability* of an even A given an event A' with nonzero probability is $\mathbb{P}(A|A') = \mathbb{P}(A \cap A')/\mathbb{P}(A')$.

We have $\mathbb{P}(A|A') = \mathbb{P}(A)$ if and only if A and A' are independent.

We can also condition random elements on an event.

Definition 2.12. Suppose $X : (\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$ is a random element, and $A_1 \in \mathcal{A}_1$ is an event with nonzero probability. Then the *distribution of X conditioned on A_1* is denoted by $\mathcal{L}_{X|A_1}$ and defined by $\mathcal{L}_{X|A_1}(A_2) = \mathbb{P}(X \in A_2|A_1)$ for $A_2 \in \mathcal{A}_2$.

Given a finite collection of measure spaces $(\Omega_1, \mathcal{A}_1, \mu_1), \ldots, (\Omega_n, \mathcal{A}_n, \mu_n)$, recall the construction of the product measure space $(\Omega, \mathcal{A}, \mu) := (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \prod_{i=1}^n \mu_i)$ (see [Rud66, Chapter 8]). If a random element takes values in a product space then each component is measurable, and conversely if the components of a random tuple are measurable then that tuple is measurable in the product space. So, we can make the following definitions:

Definition 2.13. Given random elements $X_1, \ldots, X_n$ on the same underlying probability space, $\mathcal{L}(X_1, \ldots, X_n) := \mathcal{L}((X_1, \ldots, X_n))$ is called the *joint distribution* of $X_1, \ldots, X_n$. Conversely, given a random tuple $(X_1, \ldots, X_n)$, each $\mathcal{L}(X_i)$ is called a *marginal distribution.*

Suppose we have two distributions of random elements $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$. *Coupling* is the technique of constructing a random ordered pair (X_1, X_2) which realizes the given distributions as marginal distributions. Usually this is done by specifying the joint distribution $\mathcal{L}(X_1, X_2)$.

The idea is that coupling creates a particular kind of dependence between X_1 and X_2 that allows us to compare the two distributions. Often, we are able to make conclusions about the distributions $\mathcal{L}(X_i)$ which are independent of their specific realizations as random elements in the coupling.

2.4 Expected Value

Definition 2.14. The *expected value* of a random variable X (or its distribution) is $\mathbb{E}X = \int x \, \mathrm{d}\, \mathcal{L}_X(x)$.

Remark 2.15. For a random variable X that takes integer values, this definition is equivalent to the well-known formula $\mathbb{E}X = \sum_{x \in \mathbb{Z}} x \, \mathbb{P}(X = x)$.

Remark 2.16. If X is a random variable that can be interpreted as counting the number of objects that satisfy some property, then we can express X as a sum of indicator variables $\sum_i \mathbf{1}_{A_i}$, where A_i is the event that the ith object satisfies our property. Noting that $\mathbb{E}$ is linear, we have $\mathbb{E}X = \sum_i \mathbb{E}\,\mathbf{1}_{A_i} = \sum_i \mathbb{P}A_i$. So, in order to compute the expectation of X we just need to compute the probability that each object satisfies our required property.

If we fix a particular underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can also equivalently view expectation as a linear functional on the space of integrable functions (i.e. random variables): $\mathbb{E}f = \int f(\omega) \, \mathrm{d}\,\mathbb{P}$. Sometimes we will define a new probability space $(\Omega, \mathcal{A}, \mathbb{P}')$ on an existing measurable space. In this case we will write $\mathbb{E}_{\mathbb{P}'}$ to indicate expectation with respect to the measure $\mathbb{P}'$, to avoid ambiguity. We can also define the expectation functional of a random variable $\mathbb{E}_X := \mathbb{E}_{\mathcal{L}(X)}$, so that $\mathbb{E}_X f = \mathbb{E}f(X)$.

In fact, probability measures are uniquely determined by their expectation functional, because $\mathbb{E}_{\mathbb{P}}\,\mathbf{1}_A = \mathbb{P}(A)$ for all events A. In fact, it will be an important fact for later that much weaker classes than $\{\mathbf{1}_A : A \text{ is an event}\}$ can distinguish expectation operators.

Definition 2.17. A set of real functions $\mathcal{H}$ is a *determining class* if $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$ implies that $\mathbb{P}_1 = \mathbb{P}_2$.

Another important concept later will be the idea of conditional expectation.

Definition 2.18. The expected value of a random variable with distribution $\mathcal{L}_{X|A_1}$ is called the *conditional expected value of X given* A_1 and is denoted $\mathbb{E}[X|A_1]$.

We can also define conditional expectation with respect to another random variable. If X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then the sets $X_2^{-1}(B)$ for Borel B comprise a sub-σ-algebra $\mathcal{A}'$ of $\mathcal{A}$. Then, $\mu : A' \mapsto \mathbb{E}[X_1 \, \mathbf{1}_{A'}]$ is a signed measure on $\mathcal{A}'$ that is absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}'$. By the Radon-Nikodym theorem (see [Rud66, Theorem 6.10]) there is an $\mathcal{A}'$-measurable random variable $\mathbb{E}[X_1|X_2]$ that satisfies $\mathbb{E}[X_1 \, \mathbf{1}_{A'}] = \mathbb{E}[\mathbb{E}[X_1|X_2] \, \mathbf{1}_{A'}]$ for all A' in $\mathcal{A}'$. This random variable is almost uniquely defined: for any two choices of $\mathbb{E}[X_1|X_2]$, the probability that they differ is zero.

Definition 2.19. The random variable $\mathbb{E}[X_1|X_2]$ as defined above is called the *conditional expectation of X_1 with respect to X_2*. We can also view conditional expectation as a linear operator between functions: we define $\mathbb{E}^{X_2}$ by $X_1 \mapsto \mathbb{E}[X_1|X_2]$.

Remark 2.20. This definition generalizes the previous definition of expectation conditioned on an event: if $\omega \in A$ and $\mathbb{P}(A) > 0$ then $\mathbb{E}[X|\,\mathbf{1}_A](\omega) = \mathbb{E}[X|A]$.

Remark 2.21. Note that if X_2 is discrete then we do not need to invoke Radon-Nikodym. We can define $\mathbb{E}[X_1|X_2]$ by $\mathbb{E}[X_1|X_2](\omega) = \mathbb{E}[X_1|X_2 = X_2(\omega)]$ for all $\omega \in \Omega$ with $\mathbb{P}(X_2 = X_2(\omega)) > 0$; this defines $\mathbb{E}[X_1|X_2]$ up to a set of probability zero.

We finally present a simple consequence of the definition of conditional expectation.

Proposition 2.22 (Tower Law of Expectation)**.** *Suppose X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A})$. Then $\mathbb{E}\left[\mathbb{E}^{X_2} X_1\right] = \mathbb{E}[X_1]$.*

Proof. $\mathbb{E}\left[\mathbb{E}^{X_2} X_1\right] = \mathbb{E}[\mathbb{E}[X_1|X_2]\,\mathbf{1}_\Omega] = \mathbb{E}[X_1\,\mathbf{1}_\Omega] = \mathbb{E}[X_1]$ □

2.5 Markov Chains

Comment 2.1. I'll need to define Markov Chains, stationary distributions, irreducibility and time-reversibility.

Perhaps I should talk more generally about stochastic processes, because applying exchangeable pairs to Stein's method is has connections with Ornstein-Uhlenbeck processes and also Stein's method can be applied to Poisson processes.

Issue 2.2. Every regular graph naturally defines a reversible markov chain on its vertex set, stationary with respect to the uniform distribution.

2.6 The Weak Topology on Probability Measures

Definition 2.23. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say X_n *converges in distribution* to a random variable X if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f. Alternatively, we say $\mathcal{L}(X_n)$ converges *weakly* to $\mathcal{L}(X)$, or simply $\mathcal{L}(X_n) \to \mathcal{L}(X)$.

This is only one of a number of equivalent and very natural definitions for convergence in distribution.

Definition 2.24. The *distribution function* F_X of a random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$.

Theorem 2.25. *The following are equivalent.*

(i) $\mathcal{L}(X_n) \to \mathcal{L}(X)$

(ii) $\liminf_{n\to\infty} \mathcal{L}_{X_n}(U) \geq \mathcal{L}_X(U)$ *for every open* $U \subseteq \mathbb{R}$

(iii) $F_{X_n}(x) \to F_X(x)$ *for all* x *where* F_X *is continuous*

(iv) *(Lévy's continuity theorem, see [Kal02, Theorem 4.3])* $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ *for all* $t \in \mathbb{R}$.

The equivalence of Conditions (i) to (iii) is (part of) a well-known and relatively elementary result called the Portmanteau Theorem ([Kal02, Theorem 3.25]).

When X and each X_n are integer random variables, then Condition (iii) reduces to the condition that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. This characterization is usually used to prove the Poisson limit theorem.

Classically, distributional convergence results are often proved by Lévy's continuity theorem. For example, this approach is usually used to prove the central limit theorem. For combinatorial applications, convergence in distribution can also be proved by the "method of moments": if X is the only random variable with the moments $\left(\mathbb{E}X^k\right)_{k\in\mathbb{N}}$, then $\mathcal{L}(X_n) \to \mathcal{L}(X)$ if $\mathbb{E}X_n^k \to \mathbb{E}X^k$. Convergence in distribution can also sometimes be inferred from stronger forms of convergence when X and all the X_n are coupled to the same underlying space.

A disadvantage of all these approaches is that they provide little information about the rate of convergence.

In functional analysis terms, note that expectation operators are bounded linear functionals on the space of real bounded continuous functions. Then, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ just means that $\mathbb{E}_{X_n} \to \mathbb{E}_X$ in the weak-star topology. The subspace corresponding to $\mathcal{P}(\mathbb{R})$ is confusingly called the *weak topology* on probability distributions.

Although $C_b(\mathbb{R})^*$ is not metrizable, the unit ball (in operator norm) of $C_b(\mathbb{R})^*$ is in fact metrizable (see [Rud73, Theorems 3.15 and 3.16]). Every expectation functional $\mathbb{E}$ has unit operator norm because $\mathbb{E}|h| \leq \mathbb{E}|1| = 1$ for h with unit uniform norm. So, the weak topology is metrizable.

Definition 2.26. Let $\mathcal{H}$ be a determining class of real measurable "test" functions that are uniformly absolutely bounded. Define $d_\mathcal{H} : \mathcal{P}(\mathbb{R})^2 \to \mathbb{R}^+$ by $d_\mathcal{H}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h\in\mathcal{H}}|\mathbb{E}_{\mathbb{P}_1}h - \mathbb{E}_{\mathbb{P}_2}h|$. For random variables X_1, X_2, we write $d_\mathcal{H}(X_1, X_2)$ instead of $d_\mathcal{H}(\mathcal{L}(X_1), \mathcal{L}(X_2))$.

Proposition 2.27. *Each $d_\mathcal{H}$ as defined above is a metric.*

Proof. Since the functions in $\mathcal{H}$ are bounded, $\mathbb{E}_{\mathbb{P}_1}h$ is well-defined for every $h \in \mathcal{H}$. Since the bound is uniform, the supremum in the definition of $d_\mathcal{H}$ is finite. It is immediate that $d_\mathcal{H}$ is non-negative, symmetric and satisfies the triangle inequality. Finally, $d_\mathcal{H}(\mathbb{P}_1, \mathbb{P}_2) = 0$ implies that $\mathbb{E}_{\mathbb{P}_1}h = \mathbb{E}_{\mathbb{P}_2}h$ for all $h \in \mathcal{H}$. Since $\mathcal{H}$ is a determining class, $\mathbb{P}_1 = \mathbb{P}_2$. □

Definition 2.28. We define some special cases of $d_\mathcal{H}$.

- If $\mathcal{H}_\text{K} = \left\{\mathbf{1}_{(-\infty,x]} : x \in \mathbb{R}\right\}$ then $d_\text{K} := d_{\mathcal{H}_\text{K}}$ is called the *Kolmogorov metric.*
- Let $\mathcal{H}_\text{BL}$ be the set of functions h that are absolutely bounded by 1 (that is, $|h(x)| \leq 1$ for all $x \in \mathbb{R}$), and have Lipschitz constant at most 1 (that is, $|h(x_1) - h(x_2)| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$). Then, $d_\text{BL} := d_{\mathcal{H}_\text{BL}}$ is called the *Bounded Lipschitz metric.*
- If $\mathcal{H}_\text{TV}$ is the set of functions $\mathbf{1}_B$ for Borel B, then $d_\text{TV} := d_{\mathcal{H}_\text{TV}}$ is called the *total variation metric.*

- If $\mathcal{H}_{\mathrm{W}}$ is the set of functions with Lipschitz constant at most 1, then $d_{\mathrm{W}} := d_{\mathcal{H}_{\mathrm{W}}}$ is called the *Wasserstein* metric. However, since $\mathcal{H}$ is not uniformly bounded, d_{W} is not strictly speaking a metric on $\mathcal{P}(\mathbb{R})$; the Wasserstein metric can only be used to compare distributions with finite first moment.

Proposition 2.29. *The "metrics" in Definition 2.28 are actually metrics.*

Proof. First, note that if $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R})$ have finite first moment then (with $X_1 \in \mathbb{P}_1$ and $X_2 \in \mathbb{P}_2$), then for all $h \in \mathcal{H}_{\mathrm{W}}$,

$$|\mathbb{E}h(X_1) - \mathbb{E}h(X_2)| \leq |(\mathbb{E}|X_1| + h(0)) - (\mathbb{E}|X_2| + h(0))| \leq \mathbb{E}|X_1| + \mathbb{E}|X_2|.$$

So,

$$d_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2) < \infty.$$

Now, we just need to check that each of $\mathcal{H}_{\mathrm{K}}$, $\mathcal{H}_{\mathrm{BL}}$, $\mathcal{H}_{\mathrm{TV}}$, $\mathcal{H}_{\mathrm{W}}$ are determining classes. Let $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R})$ satisfy $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for each h in $\mathcal{H}$.

If $\mathcal{H}$ is $\mathcal{H}_{\mathrm{K}}$ or $\mathcal{H}_{\mathrm{TV}}$ then $\mathbf{1}_{(-\infty,x]} \in \mathcal{H}$ for all $x \in \mathbb{R}$ so

$$\mathbb{P}_1((-\infty, x]) = \mathbb{E}_{\mathbb{P}_1} \mathbf{1}_{(-\infty,x]} = \mathbb{E}_{\mathbb{P}_2} \mathbf{1}_{(-\infty,x]} = \mathbb{P}_2((-\infty, x]).$$

Since $\{(-\infty, x] : x \in \mathbb{R}\}$ generates the Borel σ-algebra, $\mathbb{P}_1 = \mathbb{P}_2$. We have shown that $\mathcal{H}_{\mathrm{K}}$ and $\mathcal{H}_{\mathrm{TV}}$ are determining classes.

For $x \in \mathbb{R}$ and $0 < \varepsilon \leq 1$, let $h_{x,\varepsilon}$ be the continuous function which takes the value 1 on the set $(-\infty, x]$, takes the value 0 on the set $[x + \varepsilon, \infty)$, and is linearly interpolated in the range $[x, x + \varepsilon]$. Suppose $\mathcal{H}$ is $\mathcal{H}_{\mathrm{BL}}$ or $\mathcal{H}_{\mathrm{W}}$, so that each $\varepsilon h_{x,\varepsilon} \in \mathcal{H}$. It follows that $\mathbb{E}_{\mathbb{P}_1} h_{x,\varepsilon} = \mathbb{E}_{\mathbb{P}_2} h_{x,\varepsilon}$ for each $x \in \mathbb{R}$ and $0 < \varepsilon \leq 1$. For each $x \in \mathbb{R}$, note that $h_{x,1/n} \to \mathbf{1}_{(-\infty,x]}$ pointwise, and each $h_{x,1/n} \leq 1$. By the dominated convergence theorem (see [Rud66, Theorem 1.34]),

$$\mathbb{P}_1((-\infty, x]) = \mathbb{E}_{\mathbb{P}_1} \mathbf{1}_{(-\infty,x]} = \lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_1} h_{x,1/n} = \lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_2} h_{x,1/n} = \mathbb{E}_{\mathbb{P}_2} \mathbf{1}_{(-\infty,x]} = \mathbb{P}_2((-\infty, x]),$$

so $\mathcal{H}_{\mathrm{BL}}$ and $\mathcal{H}_{\mathrm{W}}$ are determining classes, as above. □

Proposition 2.30. *The metrics in Definition 2.28 are each stronger than the weak topology.*

Proof. We show that $d_{\mathcal{H}}(X_n, X) \to 0$ implies $\mathcal{L}(X_n) \to \mathcal{L}(X)$ for $\mathcal{H} \in \{\mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{W}}, \mathcal{H}_{\mathrm{TV}}\}$.

If $d_{\mathrm{K}}(X_n, X) \to 0$ or $d_{\mathrm{TV}}(X_n, X) \to 0$ then $F_{X_n} \to F_X$ uniformly, so certainly Condition (iii) of Theorem 2.25 holds.

Now, suppose $d_{\mathrm{BL}}(X_n, X) \to 0$ (this will automatically be true if $d_{\mathrm{K}}(X_n, X) \to 0$). Let $d_n = \sqrt{d_{\mathrm{BL}}(X_n, X)}$ and recall the definition of $h_{x,\varepsilon}$ from the proof of Proposition 2.29. Since $d_n h_{x,d_n} \in \mathcal{H}_{\mathrm{W}}$ for each $n \in \mathbb{N}$, we have

$$\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n} \leq d_{\mathrm{W}}(X_n, X)/d_n = d_n \to 0$$

uniformly for $x \in \mathbb{R}$. Now, note that

$$F_X(x - \varepsilon) \leq \mathbb{E}_X h_{x-\varepsilon,\varepsilon} \leq F_X(x) \leq \mathbb{E}_X h_{x,\varepsilon} \leq F_X(x + \varepsilon)$$

for any random variable X. If F_X is continuous at x then

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n}) + (F_X(x + d_n) - F_X(x)) \to 0 \\ F_{X_n}(x) - F_X(x) &\geq (\mathbb{E}_{X_n} h_{x-d_n,d_n} - \mathbb{E}_X h_{x-d_n,d_n}) + (F_X(x - d_n) - F_X(x)) \to 0 \end{aligned}$$

so Condition (iii) of Theorem 2.25 holds. □

Theorem 2.31. *The bounded Lipschitz metric metrizes the weak topology.*

We will need a small lemma to prove Theorem 2.31.

Lemma 2.32. *Let $S \subseteq \mathcal{P}(\mathbb{R})$ be compact in the weak topology. For each $\varepsilon > 0$, there is $k \in \mathbb{N}$ such that with $\sup_{\mathbb{P} \in S} \mathbb{P}((-k, k)^c) \leq \varepsilon$.*

Proof. Fix $\varepsilon > 0$. Suppose for the purpose of contradiction that for all $k \in \mathbb{N}$ there is $\mathbb{P}_k \in S$ with $\mathbb{P}_k((-k, k)^c) > \varepsilon$. Since S is compact, there is a subsequence $(\mathbb{P}_{k_n})_{n \in \mathbb{N}}$ and a measure

$\mathbb{P} \in S$ with $\mathbb{P}_{k_n} \to \mathbb{P}$ as $n \to \infty$. By Condition (ii) of Theorem 2.25, for each $k \in \mathbb{N}$ we have

$$\mathbb{P}((-k,k)) \leq \liminf_{n\to\infty} \mathbb{P}_{k_n}((-k,k)) \leq \liminf_{n\to\infty} \mathbb{P}_{k_n}((-k_n,k_n)) \leq 1-\varepsilon.$$

This is a contradiction because $\mathbb{P}((-k,k)) = \mathbb{E}_{\mathbb{P}}\, \mathbf{1}_{(-k,k)} \to \mathbb{E}_{\mathbb{P}} 1 = 1$ as $k \to \infty$, by the dominated convergence theorem (see [Rud66, Theorem 1.34]). □

Proof of Theorem 2.31. (Adapted from [Vil03, Theorem 7.12]). Let $\mathbb{P}_n \to \mathbb{P}_\infty$ weakly, and suppose for the purpose of contradiction that $d_{\mathrm{BL}}(\mathbb{P}_n, \mathbb{P}_\infty) \not\to 0$. Then there is $\varepsilon > 0$ and a subsequence $(\mathbb{P}_{k_n})_{n\in\mathbb{N}}$ with $d_{\mathrm{BL}}(\mathbb{P}_{k_n}, \mathbb{P}_\infty) > 2\varepsilon$ for all n. To simplify notation, redefine $\mathbb{P}_n$ to be $\mathbb{P}_{k_n}$ for each n (we still have $\mathbb{P}_n \to \mathbb{P}_\infty$ weakly).

Now, by the definition of d_{BL}, for each $n \in \mathbb{N}$ there is $h_n = h_n^{(0)} \in \mathcal{H}_{\mathrm{BL}}$ with

$$|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}} h_n| \geq d_{\mathrm{BL}}(\mathbb{P}_n, \mathbb{P}) - \varepsilon > \varepsilon.$$

For each $k \in \mathbb{Z}^+$, let $\mathcal{H}_{\mathrm{BL}}^{(k)} = \left\{ h|_{[-k,k]} : h \in \mathcal{H}_{\mathrm{BL}} \right\}$. By the Arzela-Ascoli theorem (see [Rud66, Theorem 11.28]), each $\mathcal{H}_{\mathrm{BL}}^{(k)}$ is a compact subset of $C_b([-k,k])$ with the uniform norm. So, for each $k \in \mathbb{Z}^+$, we can inductively choose a subsequence $\left(h_n^{(k)}\right)_{n\in\mathbb{N}}$ of $\left(h_n^{(k-1)}\right)_{n\in\mathbb{N}}$ such that $h_n^{(k)}|_{[-k,k]}$ converges uniformly to some $h^{(k)} \in \mathcal{H}_{\mathrm{BL}}^{(k)}$.

Note that $h_n^{(n)}|_{[-k,k]} \to h^{(k)}$ uniformly for each $k \in \mathbb{Z}^+$, so that $h_n^{(n)}$ converges pointwise to some function $h : \mathbb{R} \to \mathbb{R}$ that satisfies $h|_{[-k,k]} = h^{(k)}$ for all $k \in \mathbb{Z}^+$. Since h is bounded by 1 and has Lipschitz constant less than 1 on each $[-k,k]$, it follows that $h \in \mathcal{H}_{\mathrm{BL}}$ and in particular h is bounded and continuous. Finally, note

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}_\infty} h_n| \leq{} & \left|\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]}\right]\right| + \left|\mathbb{E}_{\mathbb{P}_\infty}\left[(h_n - h)\,\mathbf{1}_{[-k,k]}\right]\right| \\
& + \left|\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right]\right| + \left|\mathbb{E}_{\mathbb{P}_\infty}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right]\right| \\
& + |\mathbb{E}_{\mathbb{P}_n} h - \mathbb{E}_{\mathbb{P}_\infty} h|
\end{aligned}$$

By the definition of weak convergence, there is $N \in \mathbb{N}$ such that

$$|\mathbb{E}_{\mathbb{P}_n} h - \mathbb{E}_{\mathbb{P}_\infty} h| \leq \frac{\varepsilon}{5}$$

for $n > N$. Since the weak topology is metrizable, $\{\mathbb{P}_n\}_{n\in\mathbb{N}\cup\{\infty\}}$ is compact so by Lemma 2.32, there is $k \in \mathbb{N}$ such that

$$\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right] \leq 2\mathbb{P}_n[(-k,k)^c] \leq \frac{\varepsilon}{5}$$

for $n \in \mathbb{N} \cup \{\infty\}$. Since $\left(h_n^{(n)}\right)_{n\in\mathbb{N}}$ is a subsequence of $(h_n)_{n\in\mathbb{N}}$ and $h_n^{(n)}\,\mathbf{1}_{[-k,k]}$ converges uniformly to $h\,\mathbf{1}_{[-k,k]}$, there is $n > N$ such that

$$\left\|(h_n - h)\,\mathbf{1}_{[-k,k]}\right\|_\infty \leq \frac{\varepsilon}{5}.$$

For this n we have $|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}_\infty} h_n| \leq \varepsilon$. This is a contradiction. □

Proposition 2.30 tells us that our selection of “special” metrics are all “consistent” with weak convergence in some way, and Theorem 2.31 tells us that convergence in the Bounded Lipschitz metric is exactly the same as weak convergence. Typically, it will be natural to work with the total variation metric for Poisson approximation, and to work with the Wasserstein metric for Normal approximation. In applications, we may be most interested in the Kolmogorov metric. Therefore, it is sometimes useful to transfer results between metrics (though, this usually results in worse constants than working directly in the desired metric).

Issue 2.3. It may be worthwhile to actually characterize the Wasserstein, Kolmogorov and Total Variation topologies. In particular, Wasserstein convergence is just weak convergence plus convergence of the first moment.

Definition 2.33. If (for all x), $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}\,x$ for some f_X, then f_X is called the *Lebesgue density* of X, and X is called a *continuous* random variable.

If X is a continuous random variable, then $\mathbb{E}_X\,\mathbf{1}_A = \int_{\mathbb{R}} \mathbf{1}_A(x) f_X(x)\,\mathrm{d}\,x$, so by linearity $\mathbb{E}_X h = \int_{\mathbb{R}} h(x) f_X(x)\,\mathrm{d}\,x$ for simple h. This extends to all h by the measure-theoretic definition of the integral.

Proposition 2.34. *Let X_1, X_2 be random variables.*

(i) $d_{\mathrm{K}}(X_1, X_2) \le d_{\mathrm{TV}}(X_1, X_2)$

(ii) $d_{\mathrm{BL}}(X_1, X_2) \le d_{\mathrm{W}}(X_1, X_2)$

(iii) *If* $|f_{X_2}(x)| \le C$ *for all* x*, then* $d_{\mathrm{K}}(X_1, X_2) \le \sqrt{2C d_{\mathrm{BL}}(X_1, X_2)}$*.*

Proof. (Adapted from [Ros11, Proposition 1.2]). Items (i) and (ii) are immediate from the definition. Then, as in the proof of Proposition 2.30,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\le (\mathbb{E}_{X_n} h_{x,\varepsilon} - \mathbb{E}_X h_{x,\varepsilon}) + (\mathbb{E}_X h_{x,\varepsilon} - F_X(x)) \\ &\le d_{\mathrm{BL}}(X_1, X_2)/\varepsilon + \int_x^{x+\varepsilon} h_{x,\varepsilon}\, f_X(x)\,\mathrm{d}\,x \\ &\le d_{\mathrm{BL}}(X_1, X_2)/\varepsilon + C\varepsilon/2 \end{aligned}$$

and similarly

$$F_{X_n}(x) - F_X(x) \ge -d_{\mathrm{BL}}(X_1, X_2)/\varepsilon - C\varepsilon/2,$$

So, we can take $\varepsilon = \sqrt{2d_{\mathrm{BL}}(X_1, X_2)/C}$ to prove Item (iii). □

Example 2.35. If $\mathcal{L}_{X_2} = \mathcal{N}(0,1)$ then $f_{X_2}(x) = (2\pi)^{-1/2} e^{-x^2/2}$ so we can take $C = (2\pi)^{-1/2}$ to obtain $d_{\mathrm{K}} \le (2/\pi)^{1/4} \sqrt{d_{\mathrm{BL}}(X_1, X_2)}$.

3 Random Combinatorial Structures

Definition 3.1. Given a finite space of combinatorial objects Ω, a probability space $\left(\Omega, 2^\Omega, \mathbb{P}\right)$ is often called a *model* of Ω.

Definition 3.2. In a probability space $\left(\Omega, 2^\Omega, \mathbb{P}\right)$ where Ω is finite, if $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, then we say the space is *uniform.*

Uniform models are the simplest examples of random structures. For example, the uniform space $\mathcal{S}_n$ of permutations on n elements has $\mathbb{P}(\sigma) = 1/n!$ for each $\sigma \in S_n$. The uniform random graph model $\mathcal{G}_{n,M}$ has $\mathbb{P}(G) = \binom{\binom{n}{2}}{M}^{-1}$ for each graph G on the vertex set $[n]$ which has M edges. The uniform random regular graph model $\mathcal{G}_{n,d}$ is uniform on the set of all d-regular graphs on the vertex set $[n]$, though an explicit formula for the number of such graphs is not known.

As an important example of a (generally) non-uniform model, the (Erdős-Rényi) binomial random graph model $\mathcal{G}_{n,p}$ has

$$\mathbb{P}(G) = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}$$

for each graph G on the vertex set $[n]$. When $p = 1/2$, we obtain the uniform model on all graphs on the vertex set $[n]$.

One way to conceptualize the binomial model is to consider a sequence of independent coin tosses, where the coin is biased to land heads with probability p. Each coin toss corresponds to a particular potential edge, and determines whether that edge is present in the final random graph. When we define more complicated random models, we will often use this kind of informal description rather than giving an explicit formula for each $\mathbb{P}(\omega)$.

As another example, the uniform model $\mathcal{G}_{n,M}$ can be alternatively defined recursively: $\mathcal{G}_{n,0}$ is always the trivial graph with no edges, and for each $M > 0$, to obtain $\mathcal{G}_{n,M}$ we choose $G \in \mathcal{G}_{n,M-1}$ and add one of the $\binom{n}{2} - (M-1)$ possible edges at random.

Comment 3.1. This section is unfinished, I'll probably want random matrices and maybe the pairing model on random regular graphs

4 Stein's Method in Generality

Comment 4.1. There are a few quite different presentations of Stein's method. One thing I'm trying to do here is to unify Stein's functional analysis approach for exchangeable pairs [Ste86] with Ross' general presentation[Ros11].

The reason I want to look at Stein's original, more abstract presentation is that I think it does a better job motivating why things work. Before I read that, the steps taken to apply Stein's method seemed like blindly doing things and it turns out they work.

Issue 4.2. Maybe go through a bare-hands proof of Berry Esseen throughout this section

Suppose we have a potentially complicated random variable X, and we believe the distribution of X is close to a "standard" distribution $\mathcal{L}_0$. Then, Stein's method allows us to compare the operators $\mathbb{E}_X$ and $\mathbb{E}_0 := \mathbb{E}_{\mathcal{L}_0}$. This is sometimes directly useful for approximating statistics of X (for example, $\mathbb{P}(X \in A) = \mathbb{E}_X \mathbf{1}_A$). However, particularly for combinatorial applications, Stein's method is most often used to bound the distance $d_{\mathcal{H}}(\mathcal{L}_X, \mathcal{L}_0)$ for some $\mathcal{H}$, where the metric $d_{\mathcal{H}}$ from Definition 2.26 is defined in terms of $\mathbb{E}_X$ and $\mathbb{E}_0$.

Stein's method is motivated by the idea of a characterizing operator.

Definition 4.1. Let $\mathcal{F}_0$ be a vector space and $\mathcal{X}_0$ be a vector space of measurable functions which contains the constant functions. We say a linear operator $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is a *characterizing operator* for the distribution $\mathcal{L}_0$ if $\operatorname{im} T_0 = \mathcal{X}_0 \cap \ker \mathbb{E}_0$. For convenience, where there is no ambiguity we will often implicitly restrict $\mathbb{E}_0$ to $\mathcal{X}_0$, so we can write $\operatorname{im} T_0 = \ker \mathbb{E}_0$.

The following proposition shows why T_0 is called a characterizing operator.

Proposition 4.2. *If* $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ *is a characterizing operator and* $\mathcal{X}_0$ *is a determining class then* $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_X$ *implies* $\mathcal{L}_X = \mathcal{L}_0$.

Proof. If $h \in \mathcal{X}_0$, then $h - \mathbb{E}_0 h \in \ker \mathbb{E}_0 = \operatorname{im} T_0$ so $\mathbb{E}_X[h - \mathbb{E}_0 h] = 0$. That is, $\mathbb{E}_X h = \mathbb{E}_0 h$ for all $h \in \mathcal{X}_0$, which means $\mathcal{L}_X = \mathcal{L}_0$ by the definition of a determining class. □

Proposition 4.3. *$T_0 : \mathcal{F}_0 \to \mathcal{X}_0$ is characterizing if and only if there is a linear operator $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$ (called a* Stein transform*) such that the following two equations hold.*

$$\mathbb{E}_0 T_0 = 0_{\mathcal{F}_0}, \tag{4.1}$$

$$T_0 U_0 + \mathbb{E}_0 = \mathrm{id}_{\mathcal{X}_0}. \tag{4.2}$$

Proof. Suppose T_0 is a characterizing operator. Equation (4.1) is immediate. Let $\{h_i\}_{i\in\mathcal{I}}$ be a (Hamel) basis of $\mathcal{X}_0$. For each $i \in \mathcal{I}$ we have $h_i - \mathbb{E}_0 h_i \in \ker \mathbb{E}_0$ so there is some f_i (not necessarily unique) that solves $T_0 f_i = h_i - \mathbb{E}_0 h_i$. The operator U_0 can then be defined by $\sum_{i\in\mathcal{I}} a_i h_i \mapsto \sum_{i\in\mathcal{I}} a_i f_i$, satisfying (4.2).

Conversely, suppose (4.1) holds and U_0 exists satisfying (4.2). For $h \in \ker \mathbb{E}_0$ we have $T_0(U_0 h) = h$ and hence $h \in \operatorname{im} T_0$, so $\ker \mathbb{E}_0 \subseteq \operatorname{im} T_0$. Equation (4.1) immediately says that $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_0$, so T_0 is a characterizing operator. □

Remark 4.4. In full generality (and in accordance with [Ste86]), we do not require any topological structure on $\mathcal{F}_0$ and $\mathcal{X}_0$. The proof of Proposition 4.3 uses the axiom of choice and does not ensure that the Stein transform U_0 is particularly well-behaved. In practice, we will usually require that U_0 is well-behaved to apply Stein's method (in fact, we will usually have an explicit formula for U_0).

Comment 4.3. I managed to prove a necessary and sufficient condition for U_0 to be bounded if T_0 is a Banach space operator (namely, $\ker T_0$ is complementable). I think it's probably a little off-topic to include this though.

We'll use Proposition 4.3 to give two important examples of characterizing operators.

Theorem 4.5. *Let $X_{\mathcal{N}} \in \mathcal{N}(0,1)$ and let $\mathcal{X}_{\mathcal{N}} = \Big\{h \in C^\infty(\mathbb{R}) : \mathbb{E}\Big[|X_{\mathcal{N}}|^k |h(X_{\mathcal{N}})|\Big] < \infty \text{ for all } k \in \mathbb{N}\Big\}$. Let $\mathcal{F}_{\mathcal{N}}$ be the set of continuously differentiable f with $f' \in \mathcal{X}_{\mathcal{N}}$.*

The operator $T_{\mathcal{N}} : \mathcal{F}_{\mathcal{N}} \to \mathcal{X}_{\mathcal{N}}$ given by $T_{\mathcal{N}} f(x) = f'(x) - x f(x)$ is a characterizing operator for $\mathcal{N}(0,1)$.

Proof. (Adapted from [Ste86, Lecture II]). First we prove that $T_{\mathcal{N}}$ is well-defined as an operator with codomain $\mathcal{X}_{\mathcal{N}}$. Fix $f \in \mathcal{F}_{\mathcal{N}}$. We have

$$\mathbb{E}\Big[|X_{\mathcal{N}}|^k |T_{\mathcal{N}} f(X_{\mathcal{N}})|\Big] \le \mathbb{E}\Big[|X_{\mathcal{N}}|^k |f'(X_{\mathcal{N}})|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1} |f(X_{\mathcal{N}})|\Big].$$

Now, $\mathbb{E}|X_{\mathcal{N}}|^{k+1} < \infty$ (this is a standard result that is easily proved inductively, using integration by parts and the Gaussian integral). So,

$$\begin{aligned}\mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1} |f(X_{\mathcal{N}})|\Big] &\prec \int_{-\infty}^{\infty} |x|^{k+1} |f(x) - f(0)| e^{-x^2/2} \,\mathrm{d}\, x \\ &= \int_{-\infty}^{\infty} x^{k+1} \left| f(0) + \int_0^x f'(t) \,\mathrm{d}\, t \right| e^{-x^2/2} \,\mathrm{d}\, x \\ &\le \mathbb{E}|X_{\mathcal{N}}|^{k+1} \left(f(0) + \int_{-\infty}^{\infty} \big|f'(t)\big| \,\mathrm{d}\, t \right) < \infty.\end{aligned}$$

It follows that $\mathbb{E}\Big[|X_{\mathcal{N}}|^k |T_{\mathcal{N}} f(X_{\mathcal{N}})|\Big] < \infty$ and $T_{\mathcal{N}} f \in \mathcal{X}_{\mathcal{N}}$.

For any $f \in \mathcal{F}_{\mathcal{N}}$, integration by parts gives

$$\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} f'(t) \,\mathrm{d}\, t - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2/2} f(t) \,\mathrm{d}\, t = 0$$

so $\mathbb{E}_{\mathcal{N}} T_{\mathcal{N}} = 0$ and (4.1) holds. Then, define $U_{\mathcal{N}} : \mathcal{X}_{\mathcal{N}} \to \mathcal{F}_{\mathcal{N}}$ by

$$U_{\mathcal{N}} h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t = -e^{x^2/2} \int_{x}^{\infty} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t.$$

(These two definitions are the same because their difference is $e^{x^2/2}(\mathbb{E}_{\mathcal{N}} h - \mathbb{E}_{\mathcal{N}} h) = 0$).

A straightforward calculation involving the product rule and the fundamental theorem of calculus yields

$$T_{\mathcal{N}} U_{\mathcal{N}} h(x) = h(x) - \mathbb{E}_{\mathcal{N}} h.$$

(Indeed, $U_{\mathcal{N}}$ could have been obtained by solving this differential equation). So, (4.2) holds. To apply Proposition 4.3, it remains to prove that $U_{\mathcal{N}}$ is indeed well-defined as an operator with codomain $\mathcal{F}_{\mathcal{N}}$. For $h \in \mathcal{X}_{\mathcal{N}}$,

$$\mathbb{E}\Big[|X_{\mathcal{N}}|^k\big|(U_{\mathcal{N}}h)'(X_{\mathcal{N}})\big|\Big] \le \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1}|U_{\mathcal{N}}h(X_{\mathcal{N}})|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^k|h(x)|\Big] + \mathbb{E}|X_{\mathcal{N}}|^k\mathbb{E}_{\mathcal{N}}h.$$

By Fubini's theorem (see [Rud66, Theorem 8.8]),

$$\begin{aligned}
\mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1}|U_{\mathcal{N}}h(X_{\mathcal{N}})|\Big] &\prec \int_0^\infty x^{k+1}\int_x^\infty |h(t)-\mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\,\mathrm{d}\,t\,\mathrm{d}\,x \\
&\qquad - \int_{-\infty}^0 x^{k+1}\int_\infty^x |h(t)-\mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\,\mathrm{d}\,t\,\mathrm{d}\,x \\
&= \int_0^\infty |h(t)-\mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\int_0^t x^{k+1}\,\mathrm{d}\,x\,\mathrm{d}\,t \\
&\qquad - \int_{-\infty}^0 x^{k+1}\int_t^0 |h(t)-\mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\,\mathrm{d}\,t\,\mathrm{d}\,x \\
&\prec \int_{-\infty}^\infty |t|^{k+2}|h(t)-\mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\,\mathrm{d}\,x\,\mathrm{d}\,t < \infty,
\end{aligned}$$

so $U_{\mathcal{N}}h \in \mathcal{F}_{\mathcal{N}}$. □

Theorem 4.6. *Let* $X_{\mathrm{Po}(\lambda)} \in \mathrm{Po}(\lambda)$ *and let* $\mathcal{X}_{\mathcal{N}} = \mathcal{F}_{\mathcal{N}} = \Big\{h : \mathbb{N} \to \mathbb{R} : \mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)}\big|h\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big] < \infty \text{ for all } k \in \mathbb{N}\Big\}$

The operator $T_{\mathrm{Po}(\lambda)} : \mathcal{F}_{\mathcal{N}} \to \mathcal{X}_{\mathcal{N}}$ *defined by* $T_{\mathrm{Po}(\lambda)}f(k) = \lambda f(k+1) - kf(k)$ *is a characterizing operator for* $\mathrm{Po}(\lambda)$.

Issue 4.4. This proof is unfinished; I need to prove that $T_{\mathrm{Po}(\lambda)}$ and $U_{\mathrm{Po}(\lambda)}$ have appropriate codomain. See [Ste92], page 18.

Proof. For any $f \in \mathcal{F}_{\mathrm{Po}(\lambda)}$, we have

$$\mathbb{E}_{\mathrm{Po}(\lambda)}T_{\mathrm{Po}(\lambda)}f = e^{-\lambda}\sum_{i=0}^\infty \frac{\lambda^{i+1}}{i!}f(i+1) - e^{-\lambda}\sum_{i=1}^\infty \frac{\lambda^i}{(i-1)!}f(i) = 0$$

so $\mathbb{E}_{\mathrm{Po}(\lambda)}T_{\mathrm{Po}(\lambda)} = 0$ and (4.1) holds. Then, define $U_{\mathrm{Po}(\lambda)}$ by

$$U_{\mathrm{Po}(\lambda)}h(k) = \frac{(k-1)!}{\lambda^k}\sum_{i=0}^{k-1}\frac{\lambda^i}{i!}\big(h(i) - \mathbb{E}_{\mathrm{Po}(\lambda)}h\big)$$

for $k \geq 1$. Substituting and simplifying gives

$$T_{\mathrm{Po}(\lambda)} U_{\mathrm{Po}(\lambda)} h(k) = h(k) - \mathbb{E}_{\mathrm{Po}(\lambda)} h,$$

so (4.2) holds and Proposition 4.3 completes the proof. □

Note that $\mathcal{H}_{\mathrm{TV}} \subseteq \mathcal{X}_{\mathcal{N}}$, where $\mathcal{H}_{\mathrm{TV}}$ is as defined in Definition 2.28. Since $\mathcal{H}_{\mathrm{TV}}$ is a determining class, $T_{\mathcal{N}}$ is a characterizing operator in the sense of Proposition 4.2. We can say the same about $T_{\mathrm{Po}(\lambda)}$ if we restrict our attention to integer-valued random variables.

The utility of the introduction of a characterizing operator is that for each $h \in \mathcal{X}_0$, Equation (4.2) allows us to to make the transformation

$$\mathbb{E}_X h = \mathbb{E}_0 h + \mathbb{E}_X T_0 U_0 h. \tag{4.3}$$

The original purpose of Stein's method was to estimate some particular $\mathbb{E}_X h$. If $\mathcal{L}_0$ was chosen to be a "simple", well-understood distribution then the term $\mathbb{E}_0 h$ should be easy to compute or estimate, and if the distribution of X was "close" to $\mathcal{L}_0$, then it should be possible to show that the remainder $\mathbb{E}_X T_0 U_0 h$ is small.

For our purposes, the main use of (4.3) is to bound $d_{\mathcal{H}}(X, \mathcal{L}_0)$ for some $\mathcal{H} \subseteq \mathcal{X}_0$. For any $\mathcal{Y} \supseteq U_0 \mathcal{H}$, we have

$$d_{\mathcal{H}}(X, \mathcal{L}_0) = \sup_{h \in \mathcal{H}} |\mathbb{E}_X T_0 U_0 h| \leq \sup_{f \in \mathcal{Y}} |\mathbb{E}_X T_0 f|.$$

We have reduced the problem of bounding $d_{\mathcal{H}}(X, \mathcal{L}_0)$ to that of bounding $|\mathbb{E}_X T_0 f|$ (uniformly over $f \in \mathcal{Y}$). Especially in the cases where $\mathcal{L}_0$ is normal or Poisson and $\mathcal{H}$ is one of the standard choices in Definition 2.28, there are a number of known convenient choices of $\mathcal{Y}$, and a number of methods that are known to be effective to bound $|\mathbb{E}_X T_0 f|$.

Example 4.7. If $\mathcal{H} = \mathcal{H}_{\mathrm{TV}}$ and $\mathcal{L}_0 = \mathrm{Po}(\lambda)$, using the characterizing operator in Theorem 4.6, then we can choose

$$\mathcal{Y} = \Big\{ f \in \mathcal{F}_0 : \|f\|_\infty \leq \min\Big\{1, \lambda^{-1/2}\Big\},\ \|\Delta f\|_\infty \leq \min\Big\{1, \lambda^{-1}\Big\} \Big\},$$

where $\Delta f(k) = f(k+1) - f(k)$.

Issue 4.5. Maybe I can bound $\left\|U_{\mathrm{Po}(\lambda)}\right\|$ instead?

Proving that this choice of $\mathcal{Y}$ satisfies $\mathcal{Y} \supseteq U_{\mathrm{Po}(\lambda)}\mathcal{H}$ is nontrivial. But, we can prove that the constraints are of the "correct" order of magnitude.

Issue 4.6. The proof is in [BHJ92, Remark 10.2.4]. There's also a simpler proof in [BHJ92, Lemma 1.1.1] that $\|f\|_\infty \le 2\min\{1, \lambda^{-1/2}\}$ suffices. I'll revisit this later.

Proof that this $\mathcal{Y}$ satisfies $\mathcal{Y} \supseteq U_{\mathrm{Po}(\lambda)}\mathcal{H}$. (Adapted from [BHJ92, Lemma 1.1.1]). For any Borel A and any $k \in \mathbb{N}$,

$$\begin{aligned}
U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k) &= \frac{e^\lambda\,(k-1)!}{\lambda^k}(\mathrm{Po}_\lambda(A \cap [k-1]) - \mathrm{Po}_\lambda(A)\,\mathrm{Po}_\lambda([k-1])) \\
&= \frac{e^\lambda\,(k-1)!}{\lambda^k}\Big(\,\mathrm{Po}_\lambda(A \cap [k-1])\Big(1 - \mathrm{Po}_\lambda([k-1])\Big) \\
&\qquad\qquad - \Big(\,\mathrm{Po}_\lambda(A) - \mathrm{Po}_\lambda(A \cap [k-1])\Big)\,\mathrm{Po}_\lambda([k-1])\Big) \\
&= \frac{e^\lambda\,(k-1)!}{\lambda^k}\Big(\,\mathrm{Po}_\lambda(A \cap [k-1])\,\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1]) - \mathrm{Po}_\lambda(A\backslash[k-1])\,\mathrm{Po}_\lambda([k-1])\Big).
\end{aligned}$$

Note that $\mathrm{Po}_\lambda(A \cap [k-1])$ is bounded above by $\mathrm{Po}_\lambda([k-1])$ and $\mathrm{Po}_\lambda(A\backslash[k-1])$ is bounded above by $\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1])$, so

$$\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k)\right| \le \frac{e^\lambda\,(k-1)!}{\lambda^k}\,\mathrm{Po}_\lambda([k-1])\,\mathrm{Po}_\lambda(\mathbb{R}\backslash[k-1]).$$

Note that we have equality when $A = [k-1]$. If $k \asymp \lambda$ as $\lambda \to \infty$ then Stirling's approximation (unfinished...) □

Example 4.8. If $\mathcal{H} = \mathcal{H}_{\mathrm{W}}$ or $\mathcal{H} = \mathcal{H}_{\mathrm{BL}}$ and $\mathcal{L}_0 = \mathcal{N}(0,1)$, using the characterizing operator in Theorem 4.6, then we can choose

$$\mathcal{Y} = \Big\{ f \in \mathcal{F}_0 : \|f\|_\infty \leq 2,\ \|f'\|_\infty \leq \sqrt{2/\pi},\ \|f''\|_\infty \leq 4 \Big\}.$$

Proof that this $\mathcal{Y}$ satisfies $\mathcal{Y} \supseteq U_\mathcal{N}\mathcal{H}$. to do □

4.1 The Berry-Esseen Theorem

Before we proceed further, we give a proof of a quantitative central limit theorem using Stein's method bare-handed.

Issue 4.7. give some background; the theorem was apparently originally proved before Stein but the proof was complicated.

Theorem 4.9 (Berry-Esseen)**.** *Let $Q_1, \ldots, Q_n$ be independent random variables with common distribution Q, such that $\mathbb{E}Q = 0$, $\mathbb{E}Q^2 = 1$ and $\mathbb{E}|Q|^3 < \infty$. Let*

$$X = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i.$$

Then

$$d_{\mathrm{W}}\Big(n^{-1/2}X, \mathcal{N}(0,1)\Big) \leq \frac{3}{\sqrt{n}} \mathbb{E}\Big(|Q|^3\Big).$$

Proof. Fix $f \in \mathcal{Y}$.

We need to compare

$$\mathbb{E}[Xf(X)] = \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i f(X)\right]$$

with $\mathbb{E}f'(X)$. Define $W_i = X - \frac{1}{\sqrt{n}}Q_i$, so W_i and Q_i are independent. For each i, we then have

$$\mathbb{E}[Q_i f(X)] = \mathbb{E}[Q_i(f(X) - f(W_i))] = \frac{1}{\sqrt{n}}\mathbb{E}\big[Q_i^2 f'(W_i)\big] + \frac{1}{n}\mathbb{E}\Big[Q_i^3 \frac{f''(E)}{2}\Big]$$

for some (random) E, by Taylor's theorem with the Lagrange form of the remainder (noting $X - W_i = Q_i$).

Now, note that

$$\mathbb{E}\big[Q_i^2 f'(W_i)\big] = \mathbb{E}Q_i^2 \mathbb{E}f'(W_i) = \mathbb{E}f'(W_i) = \mathbb{E}f'(X) + \mathbb{E}\big[Q_i f''\big(E'\big)\big]$$

for some E', again using the Lagrange form of the remainder. Now, by the bound on $\|f''\|_\infty$ in the definition of $\mathcal{Y}$, we have

$$\begin{aligned}\mathbb{E}\left|Q_i^3 \frac{f''(E)}{2}\right| &\leq \mathbb{E}\big|Q^3\big|,\\ \mathbb{E}\big|Q_i f''\big(E'\big)\big| &\leq 2\mathbb{E}|Q|\end{aligned}$$

Hölder's inequality (see [Rud66, Theorem 3.6 (1)]) for the $L^{3/2}$ norm gives

$$1 = \mathbb{E}\big|Q^2\big|^{3/2} \leq \left(\mathbb{E}\left|\big(Q^2\big)^{3/2}\right|^{2/3}\right)^{3/2} = \mathbb{E}\big|Q^3\big|,$$

so another application of Hölder's inequality for the L^3 norm gives

$$\mathbb{E}|Q| \leq \mathbb{E}\big|Q^3\big|^{1/3} \leq \mathbb{E}\big|Q^3\big|.$$

We conclude

$$\begin{aligned}\big|\mathbb{E}[Xf(X)] - \mathbb{E}f'(X)\big| &\leq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\mathbb{E}[Q_i f(X)] - \frac{1}{\sqrt{n}}\mathbb{E}\big[Q_i^2 f'(W_i)\big]\right)\right|\\ &\quad + \left|\frac{1}{n}\sum_{i=1}^n \big(\mathbb{E}\big[Q_i^2 f'(W_i)\big] - \mathbb{E}f'(X)\big)\right|\\ &\leq \frac{1}{\sqrt{n}}\mathbb{E}\big|Q^3\big| + \frac{2}{\sqrt{n}}\mathbb{E}\big|Q^3\big|\end{aligned}$$

□

5 The method of exchangeable pairs

This is Stein's original approach, and is effective in wide generality for discrete random variables. In what follows, we assume X is discrete. Let Ω_X be the support of X.

Example 5.1 (adapted from [Ros11, Example 4.21])**.** Throughout this section, we will refer to an example problem to illustrate the principles of Stein's method for exchangeable pairs.

We will say a *fixed point* of a permutation $\sigma \in S_n$ is an index $k \in [n]$ that satisfies $\sigma(k) = k$. Let $X : S_n \to \{0\} \cup [n]$ give the number of fixed points in each permutation from S_n. We interpret X as a random variable on the underlying space $\mathcal{S}_n$.

Now, if n is large then fixed points are largely independent of each other, and each of n indices has a probability of $1/n$ to be a fixed point. So (recalling Remark 2.16), we might expect $\mathcal{L}_X$ to be "close" to $\mathrm{Po}(1)$. We will attempt to bound $d_{\mathrm{TV}}(X, \mathrm{Po}(1))$ to validate and quantify this intuition.

The general idea is that we can use an object $\mathbf{X}$ called an exchangeable pair to construct a characterizing operator $T_{\mathbf{X}}$ for X. We then use an operator α to connect the domains of $T_{\mathbf{X}}$ and T_0 in such a way that $T_{\mathbf{X}}\alpha$ approximates T_0. We then have $\mathbb{E}_X T_0 = \mathbb{E}_X(T_0 - T_{\mathbf{X}}\alpha)$, so we can use the fact that $T_0 - T_{\mathbf{X}}\alpha$ is small to bound $\mathbb{E}_X T_0 f$.

Definition 5.2. A 2-dimensional random pair $\mathbf{X} = (X_1, X_2)$ is an *exchangeable pair* if $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$. We will denote the support of $\mathbf{X}$ by $\Omega^{(2)}_{\mathbf{X}}$.

That is, a pair $\mathbf{X}$ is exchangeable if exchanging the components of the pair does not change their joint distribution. In particular, the marginal distributions of X_1 and X_2 must be the same.

Comment 5.1. All presentations of Stein's method I've seen use the notation (X, X') but I think that has the potential to be confusing because the X in that pair is defined on Ω^2 whereas the original random variable X is defined on Ω.

Proposition 5.3. *There is a natural equivalence between time-homogeneous reversible Markov chains with steady-state distribution $\mathcal{L}_X$, and exchangeable pairs with margins $\mathcal{L}_X$.*

Proof. Given an exchangeable pair $\mathbf{X}$ with margins $\mathcal{L}_X$, we can define a time-homogenous Markov chain M with transition probabilities $p(x_1, x_2) = \mathbb{P}(X_2 = x_2 | X_1 = x_1)$. With $\pi(x) = \mathbb{P}(X = x)$, we then have

$$\pi(x_1)p(x_1, x_2) = \mathbb{P}(\mathbf{X} = (x_1, x_2)) = \pi(x_2)p(x_2, x_1)$$

for any $x_1, x_2 \in \Omega_X$. So, M is reversible with steady-state distribution $\mathcal{L}_X$.

Conversely, suppose we have a time-homogeneous reversible Markov chain with steady-state distribution $\mathcal{L}_X$. Let the transition probability between x and x' be $p(x, x')$ and let the probability of state x in the steady-state distribution be $\pi(x)$. We can then define an exchangeable pair $\mathbf{X}$ by

$$\mathbb{P}(\mathbf{X} = (x_1, x_2)) = \pi(x_1)p(x_1, x_2) = \pi(x_2)p(x_2, x_1) = \mathbb{P}(\mathbf{X} = (x_2, x_1))$$

and the proposition is proved. □

Definition 5.4. We say an exchangeable pair $\mathbf{X}$ is *connected* if the corresponding Markov chain is irreducible.

Remark 5.5. We are particularly interested in exchangeable pairs $\mathbf{X}$ with marginal distributions $\mathcal{L}_{X_1} = \mathcal{L}_{X_2} = \mathcal{L}_X$. If X is defined on an underlying combinatorial probability space $\left(\Omega, 2^\Omega, \mathbb{P}\right)$, it is often convenient to first construct an exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathbb{P}$, so that the vector $\mathbf{X_W} = (X(W_1), X(W_2))$ is an exchangeable pair with margins $\mathcal{L}_X$. If (W_1, W_2) is connected, then $\mathbf{X_W}$ is connected also.

Example 5.6. We continue Example 5.1. We will define a specific exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathcal{S}_n$ by

$$\mathbb{P}((W_1, W_2) = (\sigma_1, \sigma_2)) = \begin{cases} \left(n!\binom{n}{2}\right)^{-1} & \text{if } \sigma_1 = \sigma_2(i\,j) \text{ for some transposition } (i\,j) \\ 0 & \text{otherwise.} \end{cases}$$

The relation of differing by a transposition is symmetric, so $\mathbf{W}$ is indeed an exchangeable pair. The Markov chain associated with $\mathbf{W}$ has a simple interpretation. Given a random permutation σ, to make a transition in the Markov chain we just randomly choose one of the $\binom{n}{2}$ possible transpositions and compose it with σ. Because the transpositions generate S_n, the pair $\mathbf{W}$ is connected, so we can use the construction from Remark 5.5 to produce a connected exchangeable pair $\mathbf{X}$ with margins X.

The Markov chain underlying a connected exchangeable pair can be naturally viewed as a connected one-dimensional simplicial complex. The zeroth reduced homology group $\ker \partial_0 / \operatorname{im} \partial_1$ of a connected simplicial complex has dimension zero, and this motivates the construction of a characterizing operator in a natural way. (The following theorem is self-contained and requires no knowledge of homology theory).

Theorem 5.7. *Suppose $\mathbf{X}$ is a connected exchangeable pair with margins $\mathcal{L}_X$. Let $\mathcal{F}_X \subseteq L^1(\Omega_X^2, \mathcal{L}_{\mathbf{X}})$ be the set of functions $f : \Omega_X^2 \to \mathbb{R}$ which satisfy $\mathbb{E}|f(\mathbf{X})| < \infty$, and are antisymmetric in the sense that $f(x_1, x_2) = -f(x_2, x_1)$. Let $\mathcal{X}_X = L^1(\Omega_X, \mathcal{L}_X)$ be the set of functions $h : \Omega_X \to \mathbb{R}$ that satisfy $\mathbb{E}_X|h| < \infty$.*

Define $T_{\mathbf{X}} : \mathcal{F}_X \to \mathcal{X}_X$ by $T_{\mathbf{X}} f(x) = \sum_{x_2 \in \Omega_X} f(x, x_2) p(x, x_2) = \mathbb{E}[f(\mathbf{X}) | X_1 = x]$, so that $T_{\mathbf{X}} f(X)$ is distributed as $\mathbb{E}^{X_1} f(\mathbf{X})$. Then $T_{\mathbf{X}}$ is a characterizing operator for X.

Proof. To see that $\operatorname{im} T_{\mathbf{X}} \subseteq \ker \mathbb{E}_X$, fix $f \in \mathcal{F}$ and note that by the tower law of expectation (Proposition 2.22),

$$\mathbb{E}_X T_{\mathbf{X}} f = \mathbb{E}\mathbb{E}^{X_1} f(\mathbf{X}) = \mathbb{E} f(\mathbf{X}).$$

By exchangeability and antisymmetry, $\mathbb{E} f(\mathbf{X}) = \mathbb{E} f(X_2, X_1) = -\mathbb{E} f(\mathbf{X})$, so $\mathbb{E} f(X_1, X_2) = \mathbb{E}_X T_{\mathbf{X}} f = 0$. This did not require the connectedness condition. We can similarly prove that $T_{\mathbf{X}}$ is well-defined as an operator from $\mathcal{F}_X$ to $\mathcal{X}_X$: note that $\mathbb{E}_X |T_{\mathbf{X}} f| = \mathbb{E}|f(\mathbf{X})|$ so $T_{\mathbf{X}} \mathcal{F}_X \subseteq \mathcal{X}_X$.

We will next prove $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$, but first we make some definitions. For each $x \in \Omega_X$, let h_x be the function that takes the value $\pi(x)^{-1}$ on x and is zero elsewhere, so that $h =$

$\sum_{x\in\Omega_X} h(x)\pi(x)h_x$ for each $h \in \mathcal{X}_X$. For each $(x_1, x_2) \in \Omega_{\mathbf{X}}^{(2)}$, define $f_{x_1,x_2} \in \mathcal{F}_X$ as the function that takes the value $(\pi(x_1)p(x_1, x_2))^{-1}$ on (x_1, x_2), takes the value $-(\pi(x_2)p(x_2, x_1))^{-1}$ on (x_2, x_1), and takes the value zero elsewhere. Note that this function is antisymmetric by the reversibility of the Markov chain of $\mathbf{X}$. We have $T_{\mathbf{X}} f_{x_1,x_2} = h_{x_2} - h_{x_1}$.

Let $h \in \ker \mathbb{E}_X$, and fix an arbitrary $x^* \in \Omega_X$. By the connectedness assumption, for each $x \in \Omega_X$ there is a sequence

$$x = x^{(0)}, x^{(1)}, \ldots, x^{(k-1)}, x^{(k)} = x^*$$

with $\left(x^{(i-1)}, x^{(i)}\right) \in \Omega_{\mathbf{X}}^{(2)}$ for $i = 1, \ldots k$. Note that

$$h_{x^*} - h_x = T_{\mathbf{X}} \sum_{i=1}^{k} f_{x^{(i-1)},x^{(i)}} =: T_{\mathbf{X}} f_x^*,$$

and it follows that

$$h = \sum_{x\in\Omega_X} h(x)\pi(x)(h_{x^*} - T_{\mathbf{X}} f_x^*) = (\mathbb{E}_X h)h_{x^*} - \sum_{x\in\Omega_X} T_{\mathbf{X}} h(x)\pi(x) f_x^*.$$

By assumption $(\mathbb{E}_X h) = 0$. If Ω_X is finite, as it will be in our applications, then it would immediately follow that $h \in \operatorname{im} T_{\mathbf{X}}$. Otherwise we will need some functional analysis. Note that $T_{\mathbf{X}}$ is actually an isometry between a subspace $\mathcal{F}_X$ of $L^1\left(\Omega_X^2, \mathcal{L}_{\mathbf{X}}\right)$ and $L^1(\Omega_X, \mathcal{L}_X)$. First we prove that $\mathcal{F}_X$ is closed and therefore a Banach space. For any $f \in L^1\left(\Omega_X^2, \mathcal{L}_{\mathbf{X}}\right)$, let $\bar{f}$ be defined by $(x_1, x_2) \mapsto -f(x_2, x_1)$, so that $f \in \mathcal{F}_X$ implies that $f = \bar{f}$. Suppose $f_n \to f$, with $f_n \in \mathcal{F}_X$ for all $n \in \mathbb{N}$. By exchangeability we have

$$\|f_n - f\| = \mathbb{E}|f_n(\mathbf{X}) - f(\mathbf{X})| = \mathbb{E}\left|f_n(X_2, X_1) - \bar{f}(X_2, X_1)\right| = \left\|f_n - \bar{f}\right\|$$

so $f = \bar{f}$ and $f \in \mathcal{F}_X$. Finally, it is a simple fact that an isometry between Banach spaces has a closed range. For, if $(T_{\mathbf{X}} f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\operatorname{im} T_{\mathbf{X}}$, then $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{F}_X$ which converges to some $f \in \mathcal{F}_X$. It follows that $T_{\mathbf{X}} f_n \to T_{\mathbf{X}} f \in \operatorname{im} T_{\mathbf{X}}$.

We have proved that $h \in \operatorname{im} T_{\mathbf{X}}$, completing the proof that $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$. □

The final step is to choose an operator $\alpha : \mathcal{F}_0 \to \mathcal{F}_X$ in such a way that $T_{\mathbf{X}}$ can be easily compared with $T_0\alpha$.

Comment 5.2. It's not clear if characterizing operators are in any sense unique so that for two characterizing operators T_1, T_2 there is *always* a connection α that makes $T_1 = T_2\alpha$.

Example 5.8. For the Poisson case in Theorem 4.6, we need to compare $\lambda f(X+1)$ with $Xf(X)$. It is often fruitful to define α by

$$\alpha f(x_1, x_2) = cf(x_2)\,\mathbf{1}\{x_2 = x_1 + 1\} - cf(x_1)\,\mathbf{1}\{x_1 = x_2 + 1\}$$

for some $c \in \mathbb{R}$. We will then have

$$T_0 f - T_{\mathbf{X}}\alpha f = f(X+1)(\lambda - c\mathbb{P}(X_2 = X_1 + 1|X_1)) - f(X_1)(X_1 - c\mathbb{P}(X_1 = X_2 + 1|X_1)).$$

Using the triangle inequality and the choice of $\mathcal{Y}$ in Example 4.7, for all $f \in \mathcal{Y}$:

$$\mathbb{E}_X T_0 f \le \min\Big(1, \lambda^{-1/2}\Big)(\mathbb{E}|\lambda - c\mathbb{P}(X_2 = X_1 + 1|X_1)| + \mathbb{E}|X_1 - c\mathbb{P}(X_1 = X_2 + 1|X_1)|).$$

This approximation is effective when

$$\begin{aligned} \mathbb{P}(X_2 = X_1 + 1|X_1) &\approx \lambda/c, && (5.1)\\ \mathbb{P}(X_2 = X_1 - 1|X_1) &\approx X_1/c. \end{aligned}$$

The interpretation of these approximate equalities is that the Markov chain associated with $\mathbf{X}$ is approximately an immigration-death process. This is likely to happen when $X(\omega)$ is in some sense a statistic of the amount of local structure over the object ω, and $\mathbf{X}$ is defined by a Markov chain on Ω (as in Remark 5.5) that (uniformly) randomly disturbs local structure. The conclusion to Example 5.1 should make this clear:

Issue 5.3. I should include sub-σ-algebra note to simplify conditional expectation stuff

Example 5.9. We continue Example 5.1, recalling the exchangeable pairs $\mathbf{W}$ and $\mathbf{X}$ from Example 5.6. The interpretation of

$$\mathbb{P}(X_2 = X_1 - 1|X_1)$$

is the probability of a transposition destroying exactly one out of an existing X_1 fixed points. In order to destroy exactly one fixed point, we have to choose a fixed point to destroy, and swap it with a non-fixed-point. There are $X_1(n - X_1)$ out of $\binom{n}{2}$ transpositions that do this, so

$$\mathbb{P}(X_1 = X_2 + 1|X_1) = \frac{X_1(n - X_1)}{\binom{n}{2}}.$$

Next, we will find a formula for $\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)$, noting that

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \mathbb{E}[\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)|X_1].$$

In order to create exactly one fixed point, we have to choose an index k that is not fixed in W_1 (there are $n - X_1$ such) and compose W_1 with $(k\,\sigma(k))$. This creates exactly one fixed point unless $\sigma^{-1}(k) = k$, in which case it creates two. We have counted this second case twice for every transposition in the cycle decomposition of W_1. Let Y be the number of transpositions in the cycle decomposition of W_1. We have

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{\binom{n}{2}}.$$

In order to satisfy (5.1) as closely as possible, we choose $c = \binom{n}{2}/n$. Recalling that $\mathbb{E}X_1 = 1$, we then have

$$\begin{aligned}\mathbb{E}|1 - c\mathbb{P}(X_2 = X_1 + 1|X_1)| &= \mathbb{E}\left[1 - \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{n}\right] \\ &= 1/n + 2\mathbb{E}Y/n, \\ \mathbb{E}|X_1 - c\mathbb{P}(X_2 = X_1 - 1|X_1)| &= \mathbb{E}\left[X_1 - \frac{X_1(n - X_1)}{n}\right] \\ &= \mathbb{E}\left[X_1^2\right]/n.\end{aligned}$$

Now, the probability that a transposition $(i\,j)$ is in the cycle decomposition of W_1 is $(n(n-1))^{-1}$ because i must map to j out of the n possible options in $[n]$, then j must map to i out of the $n-1$ possible options in $[n]\backslash\{j\}$. There are $\binom{n}{2} = n(n-1)/2$ possible transpositions so by Remark 2.16 it follows that $\mathbb{E}Y = 1/2$.

Now, $\mathbb{E}[X_1(X_1 - 1)/2]$ is the expected number of unordered pairs of distinct fixed points in a permutation $\sigma \in \mathcal{S}_n$. For any unordered pair of distinct indices $\{i, j\}$, the probability that both are fixed is $(n(n-1))^{-1}$ because i must map to i out of the n possible options in $[n]$, then j must map to j out of the $n-1$ possible options in $[n]\backslash\{i\}$. The total number of unordered pairs is $\binom{n}{2} = n(n-1)/2$, so again applying Remark 2.16, we have $\mathbb{E}[X_1(X_1 - 1)/2] = 1/2$ and $\mathbb{E}\left[X_1^2\right] = 2$.

We conclude that $d_{\mathrm{TV}}(X, \mathrm{Po}(1)) \leq 4/n$.

Comment 5.4. The “generator method” [BC05] says that the Poisson characterizing operator can be obtained with the generator of an immigration-death process and the Normal characterizing operator can be obtained with the generator of an O-U process. Investigate the link here?

5.1 An Application: Switchings and Short Cycles in Random Regular Graphs

A particularly interesting (and new!) application of Stein's method is to piggyback onto results proved using combinatorial *switchings.* In this subsection, I'll describe how a certain Poisson limit theorem concerning short cycles in random regular graphs was improved by applying Stein's method in a quite natural, and generally applicable, way. First, we need some background on short cycles in random regular graphs, and switchings.

5.1.1 Short Cycles in Random Regular Graphs

Let $X_{\ell,n}^{(d)}$ be the number of cycles of length ℓ in a random d-regular graph on n vertices. The following is a critical theorem describing the structure of random regular graphs.

Theorem 5.10. *Fix $\ell \in \mathbb{N}$ and $d \in \mathbb{N}$ with $d \geq 3$. Then*

$$\mathcal{L}\Big(X_{\ell,n}^{(d)}\Big) \to \mathrm{Po}\Bigg(\frac{(d-1)^{\ell}}{2\ell}\Bigg)$$

as $n \to \infty$, with n restricted to the even integers if d is odd.

One reason Theorem 5.10 is interesting is because it tells us the number of "short cycles" in a random regular graph does not grow (to infinity) with the size of the graph. That is, random regular graphs are likely to "locally" "look like forests". Theorem 5.10 is most easily proved with the method of moments; see [Wor99].

Theorem 5.10 theorem can be extended to the case where d is allowed to grow modestly with n. In [MWW04], a natural variant of Theorem 5.10 was proved for $(d-1)^{2\ell-1} = o(n)$, using the idea of *switchings.* The condition $(d-1)^{2\ell-1} = o(n)$ appeared to be a natural threshold; it was conjectured that short cycle counts are no longer asymptotically Poisson if d grows any faster.

It was recently proved that in fact cycle counts remain asymptotically Poisson past this boundary:

Theorem 5.11 ([Joh11])**.** *Let ℓ and d depend on n, in such a way that $d \geq 3$ and*

$$\sqrt{\ell}(d-1)^{3\ell/2-1} = o(n).$$

Then,

$$d_{\mathrm{TV}}\left(X_{\ell,n}^{(d)}, \mathrm{Po}\left(\frac{(d-1)^{\ell}}{2\ell} \right) \right) \to 0.$$

if $n \to \infty$ in such a way that nd is always even.

The proof of Theorem 5.11 combines switchings with Stein's method.

5.1.2 Switchings

Formally speaking, a switching is a binary relation $\rightsquigarrow$ on a set of combinatorial objects Ω (in our case, Ω is the set of d-regular graphs on n vertices). Usually, switchings are understood as an "action" that changes one combinatorial object to another. The switching used in [MWW04] is defined as follows:

Definition 5.12. Let $C = v_1 \ldots v_\ell$ be an ℓ-cycle in a d-regular graph G. For each $i \in \mathbb{Z}/\ell\mathbb{Z}$, suppose $e_i = u_i w_{i+1} \in E(G)$ satisfies $u_i v_i \notin E(G)$ and $v_i w_i \notin E(G)$. Let G' be the d-regular graph obtained from G by deleting all the edges in C and adding each of the edges $u_i v_i$ and $v_i w_i$. Further, suppose this operation does not create or delete any ℓ-cycle except C. Then, we say $G \rightsquigarrow G'$.

Issue 5.5. Need picture here. Also need precise definition: are edges/cycle oriented? Can we have some $u_i = v_j$?

The typical application of a switching is to estimate the relative sizes of some subsets of Ω, by analysing the "flow" of switchings between the subsets. For example, define $S(k)$ to be the set of d-regular graphs with k cycles of length ℓ. In [MWW04, Section 3], bounds were obtained for the "number of ways to switch out" of each graph G (that is, the number of graphs

G' satisfying $G \rightsquigarrow G'$), and the "number of ways to switch in" to each graph. This gives an estimate on the relative sizes of $S(k)$ and $S(k-1)$ for each k, which is then used to estimate the distribution of $X_{\ell,n}^{(d)}$. The relevant estimates work in the regime $(d-1)^{2\ell-1} = o(n)$.

5.1.3 Stein's method and Switchings

The application of switchings in the proof of Theorem 5.11 is different conceptually, but the same kind of estimates are needed. Hopefully it is possible that Stein's method can be applied similarly to a variety of different switching arguments unrelated to short cycles in random regular graphs.

The approach is conceptually similar to that of Example 5.1. In the case of permuations, the transposition of two random indices is a simple and effective way to perturb a permutation in an appropriate way to apply the method of Example 5.8. However, it is not so easy to perturb a regular graph in a way that maintains regularity. We accomplish this with the switching from Definition 5.12.

We define a multigraph $\mathfrak{G}$ with vertex set Ω. Make an edge in $\mathfrak{G}$ for every pair (G, G') with $G \rightsquigarrow G'$. Let Δ be the maximum degree so far. Add enough loops to each vertex of $\mathfrak{G}$ so that $\mathfrak{G}$ is a Δ-regular multigraph. Recalling Issue 2.2, $\mathfrak{G}$ induces a reversible Markov chain on Ω with stationary distribution $\mathcal{G}_{n,d}$, which corresponds to an exchangeable pair $\mathbf{G} = (G_1, G_2)$ with margins $\mathcal{G}_{n,d}$. With Remark 5.5, this provides us with an exchangeable pair $\mathbf{X} = (X_1, X_2)$ with margins $\mathcal{L}_X$ to apply Stein's method.

5.2 Proof of Theorem 5.11

(Simplified from the proof of [Joh11, Theorem 11]).

Let $X = X_{\ell,n}^{(d)}$, let Ω be the set of all d-regular graphs on n vertices, and let $\lambda = (d-1)^{\ell}/(2\ell) \geq 1$.

The approach is conceptually similar to that of Example 5.1. In the case of permuations, the transposition of two random indices is a simple and effective way to perturb a permutation in an appropriate way to apply the method of Example 5.8. However, it is not so easy to perturb a regular graph in a way that maintains regularity. We accomplish this with the switching from Definition 5.12.

We define a multigraph $\mathfrak{G}$ with vertex set Ω. Make an edge in $\mathfrak{G}$ for every pair (G, G') with $G \rightsquigarrow G'$. Let Δ be the maximum degree so far. Add enough loops to each vertex of $\mathfrak{G}$ so that $\mathfrak{G}$ is a Δ-regular multigraph. Recalling Issue 2.2, $\mathfrak{G}$ induces a reversible Markov chain on Ω with stationary distribution $\mathcal{G}_{n,d}$, which corresponds to an exchangeable pair $\mathbf{G} = (G_1, G_2)$ with margins $\mathcal{G}_{n,d}$. With Remark 5.5, this provides us with an exchangeable pair $\mathbf{X} = (X_1, X_2)$ with margins $\mathcal{L}\left(X_{\ell,n}^{(d)}\right)$ to apply Stein's method.

For $G \in \Omega$ and an ℓ-cycle C, let $F_C(G)$ be the number of ways to switch from G by destroying C. That is,

$$F_C(G) = \left|\left\{G' \in \Omega : G \rightsquigarrow G',\, C \subseteq G,\, C \nsubseteq G'\right\}\right|,$$

so that $\sum_{C \in K_n} F_C(G)$ is the number of ways to switch from G, destroying one ℓ-cycle. Then,

$$\begin{aligned}\mathbb{P}(X_2 = X_1 + 1|X_1) &= \mathbb{E}[\mathbb{P}(X_2 = X_1 + 1|G_1)|X_1] \\ &= \mathbb{E}\left[\frac{1}{\Delta}\sum_{C \in K_n} F_C(G_1)\middle| X_1\right].\end{aligned}$$

Issue 5.6. Conditioning on X_1 is annoying, I'd rather condition on G_1. I should add a remark to this effect when introducing the method in the theory section, and make the same adjustment to the permutations example.

Next, the automorphism group of an ℓ-cycle has size 2ℓ, so the number of ℓ-cycles in K_n is $(n)_\ell/(2\ell)$. It follows that

$$\begin{aligned}\mathbb{E}\left|X_1 - \frac{\Delta}{(n)_\ell d^\ell}\mathbb{P}(X_2 = X_1 + 1|X_1)\right| &= \mathbb{E}\left|\sum_{C \in K_n} \mathbf{1}_{C \subseteq G_1} - \frac{1}{(n)_\ell d^\ell}\sum_{C \in K_n} F_C(G)\right| \\ &\leq \frac{(n)_\ell}{2\ell}\mathbb{E}\left|\mathbf{1}_{C \subseteq G_1} - \frac{F_C(G_1)}{(n)_\ell d^\ell}\right|.\end{aligned}$$

Similarly, let $B_C(G)$ be the number of ways to switch into G by destroying C. We have

$$\mathbb{E}\left|\lambda - \frac{\Delta}{(n)_\ell d^\ell}\mathbb{P}(X_2 = X_1 + 1|X_1)\right| \leq \frac{(n)_\ell}{2\ell}\mathbb{E}\left|\frac{(d-1)^n}{(n)_\ell} - \frac{B_C(G_1)}{(n)_\ell d^\ell}\right|.$$

The majority of this section will consist of combinatorially bounding these expectations, resulting in the following lemmas:

Lemma 5.13. *Uniformly in C,*

$$\mathbb{E}\left|\mathbf{1}_{C\subseteq G_1} - \frac{F_C(G_1)}{(n)_\ell d^\ell}\right| = O\left(\frac{\ell(d-1)^{2\ell-1}}{n^{\ell+1}}\right).$$

Lemma 5.14. *Uniformly in C,*

$$\mathbb{E}\left|\frac{(d-1)^n}{(n)_\ell} - \frac{B_C(G_1)}{(n)_\ell d^\ell}\right| = O\left(\frac{\ell(d-1)^{2\ell-1}}{n(n)_\ell}\right).$$

After proving these lemmas, we can conclude that

$$\begin{aligned} d_{\mathrm{TV}}(X, \mathrm{Po}(\lambda)) &\leq \lambda^{-1/2}(\mathbb{E}|\lambda - \Delta\mathbb{P}(X_2 = X_1 + 1|X_1)| + \mathbb{E}|X_1 - \Delta\mathbb{P}(X_2 = X_1 + 1|X_1)|) \\ &= O\left(\frac{\sqrt{\ell}(d-1)^{3\ell/2-1}}{n}\right). \end{aligned}$$

This concludes the proof of Theorem 5.11.

In order to prove Lemma 5.13 and **??**, we will need a few lemmas.

Lemma 5.15. *Uniformly in C and G such that C does not share an edge with another ℓ-cycle in G,*

$$F_C(G) \geq (n)_\ell d^\ell\left(1 - \frac{2\ell\gamma(G) + O\left(\ell(d-1)^\ell\right)}{nd}\right),$$

where $\gamma(G)$ is the number of ℓ-cycles in G.

Proof. adsf □

Lemma 5.16. *Uniformly in C and G,*

$$B_C(G) \geq (d(d-1))^{\ell}\left(1 - \frac{O\Big(\ell(d-1)^{\ell-1}\Big)}{n}\right)$$

Proof. adsf □

Proof of Lemma 5.13. We partition $\mathcal{G}_{n,d}$ into three events:

$$\begin{aligned}
&A_1 = \{G \in \Omega : C \nsubseteq G\},\\
&A_2 = \{G \in \Omega : C \subseteq G,\, C \text{ does not share an edge with another } \ell\text{-cycle in } G\},\\
&A_3 = \{G \in \Omega : C \subseteq G,\, C \text{ shares an edge with another } \ell\text{-cycle in } G\}.
\end{aligned}$$

If $G \in A_1$ then $C \nsubseteq G$, and no switching can destroy C, so $F_C(G) = 0$. If $G \in A_3$, then a switching that destroys C also destroys some other ℓ-cycle, which is not allowed in the definition of $\rightsquigarrow$. So, $F_C(G) = 0$. Also, note that a switching that deletes C is determined by the ℓ oriented edges e_i in the definition of $\rightsquigarrow$. There are at most $(n)_\ell d^\ell$ ways to choose these edges (choose the ℓ vertices u_i then choose a neighbor w_i for each), so $F_C \leq (n)_\ell d^\ell$.

It follows that

$$\begin{aligned}
\mathbb{E}\left|\mathbf{1}_{C\subseteq G_1} - \frac{F_C(G_1)}{(n)_\ell d^\ell}\right| &= \mathbb{E}\left[\mathbf{1}_{A_1}|0-0| + \mathbf{1}_{A_2}\left|1 - \frac{F_C(G_1)}{(n)_\ell d^\ell}\right| + \mathbf{1}_{A_1}(1-0)\right]\\
&= \mathbb{E}\left[\mathbf{1}_{A_2}\left(1 - \frac{F_C(G_1)}{(n)_\ell d^\ell}\right)\right] + \mathbb{P}(A_3)
\end{aligned}$$

Note that a switching that deletes C is determined by the ℓ oriented edges e_i in the definition of $\rightsquigarrow$. There are at most $(n)_\ell d^\ell$ ways to choose these edges (choose the ℓ vertices u_i then choose a neighbor w_i for each), so $F_C \leq (n)_\ell d^\ell$.

, so $F_C(G) \leq (n)_\ell d^\ell$ and □

Proof of Lemma 5.14. asfd □

6 Size-Bias Coupling

References

[BC05] Andrew D Barbour and Louis Hsiao Yun Chen, *An introduction to Stein's method*, Lecture notes series, Institute for Mathematical Sciences, National University of Singapore, vol. 4, Singapore University Press and World Scientific, 2005.

[BHJ92] Andrew D Barbour, Lars Holst, and Svante Janson, *Poisson approximation*, Clarendon Press Oxford, 1992.

[Joh11] Tobias Johnson, *Exchangeable pairs, switchings, and random regular graphs*, arXiv preprint arXiv:1112.0704 (2011).

[Kal02] Olav Kallenberg, *Foundations of modern probability*, springer, 2002.

[MWW04] Brendan D McKay, Nicholas C Wormald, and Beata Wysocka, *Short cycles in random regular graphs*, Electron. J. Combin **11** (2004), no. 1, 1–12.

[Ros11] Nathan Ross, *Fundamentals of Stein's method*, Probab. Surv **8** (2011), 210–293.

[Rud66] Walter Rudin, *Real and complex analysis*, McGraw-Hill, 1966.

[Rud73] ______, *Functional analysis*, McGraw-Hill, 1973.

[Ste86] Charles Stein, *Approximate computation of expectations*, IMS Lecture Notes – Monograph Series, no. 7, IMS, 1986.

[Ste92] ______, *A way of using auxiliary randomization*, Probability Theory (1992), 159–180.

[Vil03] Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, no. 58, American Mathematical Society, 2003.

[Wor99] Nicholas C Wormald, *Models of random regular graphs*, London Mathematical Society Lecture Note Series (1999), 239–298.